IPW

Atty. Dkt. No. 089367-0127

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant: Fumitoshi MIZUTANI, et al.

Title: DATA PROCESSING APPARATUS AND DATA PROCESSING METHOD

Appl. No.: 10/827,433

Filing Date: 04/20/2004

Examiner: Unknown

Art Unit: Unknown

CLAIM FOR CONVENTION PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

The benefit of the filing date of the following prior foreign application filed in the following foreign country is hereby requested, and the right of priority provided in 35 U.S.C. § 119 is hereby claimed.

In support of this claim, filed herewith is a certified copy of said original foreign application:

Japanese Patent Application No. 2003-115621
filed 04/21/2003.

Respectfully submitted,

RN 38,077

Date: May 26, 2004

By _____

for David A. Blumenthal
Attorney for Applicant
Registration No. 26,257

FOLEY & LARDNER LLP
Customer Number: 22428
Telephone: (202) 672-5407
Facsimile: (202) 672-5399

US

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: ２００３年 ４月２１日

出 願 番 号
Application Number: 特願２００３－１１５６２１

[ST. 10/C]: [ＪＰ２００３－１１５６２１]

出 願 人
Applicant(s): 日本電気株式会社

２００４年 ３月 ２日

特許庁長官
Commissioner,
Japan Patent Office 今 井 康 夫

出証番号 出証特２００４－３０１５７０２

【書類名】　特許願
【整理番号】　62703096
【提出日】　平成15年 4月21日
【あて先】　特許庁長官　殿
【国際特許分類】　G06F 15/16
　　G06F 11/16
【発明者】
　【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
　【氏名】　水谷　文俊
【発明者】
　【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
　【氏名】　尾田　眞也
【特許出願人】
　【識別番号】　000004237
　【氏名又は名称】　日本電気株式会社
【代理人】
　【識別番号】　100095407
　【弁理士】
　【氏名又は名称】　木村　満
【手数料の表示】
　【予納台帳番号】　038380
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　9715824
【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　データ処理装置

【特許請求の範囲】

【請求項１】

同一のデータ送信元から同一のデータを受信する複数の受信インタフェース部を備え、前記複数の受信インタフェース部が受信したデータの処理を並列して行うデータ処理装置において、

前記各受信インタフェース部は、受信したデータにエラーが発生すると、データの受信を停止し、異なる受信インタフェース部に、前記データ送信元からのデータ受信を停止させる通信エラー信号を出力して、前記データ送信元にデータの再送を要求する通信エラー処理部を備えた、

ことを特徴とするデータ処理装置。

【請求項２】

前記各受信インタフェース部の通信エラー処理部は、受信したデータの一部にエラーが発生すると、エラーが発生したデータを破棄して、前記データ送信元に、破棄したデータの再送を要求するように構成されたものである、

ことを特徴とする請求項１に記載のデータ処理装置。

【請求項３】

前記データ送信元は、同一のシリアルデータを送信するものであって、

前記各受信インタフェース部の通信エラー処理部は、受信したシリアルデータにエラーが発生すると、エラーが発生したシリアルデータ及び当該データに続いて受信したシリアルデータを破棄して、前記データ送信元に破棄したシリアルデータの再送を要求するように構成された、

ことを特徴とする請求項１に記載のデータ処理装置。

【請求項４】

前記データ送信元は、各パケットにシーケンス番号を付加してパケット単位で前記データを送信するものであって、

前記各受信インタフェース部の通信エラー処理部は、受信したパケットのデータにエラーが発生すると、前記受信した各パケットに付加されたシーケンス番号

に基づいてパケット単位で、前記データ送信元に、データの再送を要求するように構成されたものである、

ことを特徴とする請求項１乃至３のいずれか１項に記載のデータ処理装置。

【請求項５】

所定のクロック信号の周波数を分周して同期信号を生成し、生成した同期信号を前記各受信インタフェース部に供給する分周器を備え、

前記各受信インタフェース部は、前記分周器が供給した同期信号に従ってデータを受信するものである、

ことを特徴とする請求項１乃至４のいずれか１項に記載のデータ処理装置。

【請求項６】

送信対象のデータを同じタイミングで複数のデータ送信先に送信する送信インタフェース部を備えたデータ処理装置において、

前記送信インタフェース部は、送信対象のデータを、所定のクロック信号の１周期以内で送信可能なデータ長のデータに分割してパケットデータを生成し、生成した各パケットデータを前記クロック信号に同期させて同じタイミングで前記複数の送信先に送信するように構成された、

ことを特徴とするデータ処理装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、同一のデータを並列して処理するデータ処理装置に関する。

【０００２】

【従来の技術】

データ処理を実行するコンピュータシステムとして、既存のコンポーネントを流用して、冗長化された構成のフォールトトレラントコンピュータシステムがある（例えば、特許文献１参照）。このコンピュータシステムでは、ロックステップ方式が採用されている。

【０００３】

このロックステップ方式とは、冗長化構成の複数のプロセッサを備え、複数の

プロセッサが同一のデータを同期して並列処理し、複数のプロセッサからの出力を比較してエラーを検出すると、そのエラーを修復するようにした方式である。

【０００４】

また、近年のコンピュータシステムにおいては、プロセッサとＩ／Ｏシステムとの間の接続には、PCI-Express、Hyper-Transport（登録商標）、InfiniBand（登録商標）等のような高速でデータの送受信を行える高速シリアルリンク方式が採用されつつある。

【０００５】

【特許文献１】

特開平９－１２８３４９号公報（第５－７頁、図１）

【０００６】

【発明が解決しようとする課題】

しかし、従来の冗長構成のコンピュータシステムにおいて、このような高速のデータ送受信方式を採用することにより、データの送受信速度が速くなると、複数のプロセッサが処理するデータの同一性を保証することが難しくなり、通信エラーも生じやすくなる。

【０００７】

プロセッサとＩ／Ｏシステムとの間でデータを送受信するインタフェース部は、通信エラーを検出すると、それぞれ、異なるタイミングでデータの再送信を要求してしまう。各インタフェース部が、データの再送信を要求すると、各プロセッサでの処理のタイミングや順序にずれが生じ、ロックステップ方式を維持できなくなり、複数のプロセッサが同一のデータ処理を同期して行うことが難しくなる。

【０００８】

また、このようなコンピュータシステムにおいては、通信線の配線長によってデータ遅延が生じやすい。データ遅延が生じて複数のプロセッサの処理タイミングがずれると、同じように複数のプロセッサが同一のデータ処理を同期して行うことが難しくなる。このため、等長配線を厳密に行わなければならず、システムの筐体構造やボード設計、ボード構成の自由度にも大きな制約が生じることにな

る。

【０００９】

本発明は、このような従来の問題点に鑑みてなされたもので、通信エラーが生じた場合でも同一のデータ処理を同期して行うことが可能なデータ処理装置を提供することを目的とする。

【００１０】

【課題を解決するための手段】

この目的を達成するため、本発明の第１の観点に係るデータ処理装置は、

同一のデータ送信元から同一のデータを受信する複数の受信インタフェース部を備え、前記複数の受信インタフェース部が受信したデータの処理を並列して行うデータ処理装置において、

前記各受信インタフェース部は、受信したデータにエラーが発生すると、データの受信を停止し、異なる受信インタフェース部に、前記データ送信元からのデータ受信を停止させる通信エラー信号を出力して、前記データ送信元にデータの再送を要求する通信エラー処理部を備えたものである。

【００１１】

このような構成によれば、通信エラーが生じた場合でも同一のデータ処理を同期して行うことが可能になる。

【００１２】

前記各受信インタフェース部の通信エラー処理部は、受信したデータの一部にエラーが発生すると、エラーが発生したデータを破棄して、前記データ送信元に、破棄したデータの再送を要求するように構成されたものであってもよい。

【００１３】

前記データ送信元は、同一のシリアルデータを送信するものであって、

前記各受信インタフェース部の通信エラー処理部は、受信したシリアルデータにエラーが発生すると、エラーが発生したシリアルデータ及び当該データに続いて受信したシリアルデータを破棄して、前記データ送信元に破棄したシリアルデータの再送を要求するように構成されたものであってもよい。

【００１４】

前記データ送信元は、各パケットにシーケンス番号を付加してパケット単位で前記データを送信するものであって、

前記各受信インタフェース部の通信エラー処理部は、受信したパケットのデータにエラーが発生すると、前記受信した各パケットに付加されたシーケンス番号に基づいてパケット単位で、前記データ送信元に、データの再送を要求するように構成されたものであってもよい。

【００１５】

所定のクロック信号の周波数を分周して同期信号を生成し、生成した同期信号を前記各受信インタフェース部に供給する分周器を備え、

前記各受信インタフェース部は、前記分周器が供給した同期信号に従ってデータを受信するものであってもよい。

【００１６】

本発明の第２の観点に係るデータ処理装置は、

送信対象のデータを同じタイミングで複数のデータ送信先に送信する送信インタフェース部を備えたデータ処理装置において、

前記送信インタフェース部は、送信対象のデータを、所定のクロック信号の１周期以内で送信可能なデータ長のデータに分割してパケットデータを生成し、生成した各パケットデータを前記クロック信号に同期させて同じタイミングで前記複数の送信先に送信するように構成されたものである。

【００１７】

【発明の実施の形態】

以下、本発明の実施の形態に係るデータ処理装置を図面を参照して説明する。

尚、本実施の形態に係るデータ処理装置を、冗長構成を有するコンピュータシステムとして説明する。

本実施の形態に係るコンピュータシステムの構成を図１に示す。

本実施の形態に係るコンピュータシステムは、冗長化構成の複数のプロセッサを備え、複数のプロセッサが同一のデータを同期して並列処理するロックステップ方式に従って動作するフォールトトレラントコンピュータシステムであり、サブシステム１、２を備えて構成される。

【００１８】

サブシステム１は、演算システム１１と、Ｉ／Ｏシステム１２と、を備えて構成される。サブシステム２は、演算システム２１と、Ｉ／Ｏシステム２２と、を備えて構成される。

【００１９】

演算システム１１，２１には、同期した周波数１６６ＭＨｚのクロック信号ＣＬＫが供給される。このように、演算システム１１，２１に、同期したクロック信号ＣＬＫが供給されることにより、サブシステム１，２は、ロックステップ方式に従って、同一の処理を同期して同時に実行する。

【００２０】

サブシステム１，２の間には、分周器３１が接続される。分周器３１は、ＦＳＢのクロック信号ＣＬＫが供給されて、このクロック信号ＣＬＫを分周するものであり、クロック信号ＣＬＫを分周して同期信号Ｓ１を生成する。

【００２１】

分周器３１は、生成した同期信号Ｓ１を演算システム１１のメモリブリッジ１６、演算システム２１のメモリブリッジ２６、Ｉ／Ｏシステム１２のＩ／Ｏブリッジ１８、Ｉ／Ｏシステム２２のＩ／Ｏブリッジ２８に、それぞれ、供給する。

【００２２】

尚、本実施の形態では、データの送受信に、PCI-Expressインタフェースを用いるものとする。PCI-Expressインタフェースでは、パラレルバスで生ずる信号線と信号線との間でのデータのずれ（スキュー）を防止するため、シリアルリンクが採用されている。演算システム１１，２１とＩ／Ｏシステム１２，２２との間もPCI-Expressインタフェースに従って接続される。

【００２３】

分周器３１は、同期信号Ｓ１の１周期が、2.5Gbps/laneのPCI-Expressインタフェースの２４シンボルタイムに相当するように、クロック信号ＣＬＫの周波数１６６ＭＨｚを１／１６の周波数１０．４ＭＨｚに分周する。

【００２４】

PCI-Expressインタフェースでは、演算システム１１，２１、Ｉ／Ｏシステム

１２，２２といったデバイスは、１対１で接続される。差動信号を用いてデータを伝送する場合、リンクには、片方向に２本、双方向で合計４本の信号線が用いられる。この４本の信号線の組は、レーンと呼ばれるものである。

【００２５】

また、１シンボルタイムとは、PCI-Expressに従って、１レーンのデータを８Ｂ／１０Ｂエンコーディング処理した後の１バイトの有効データを送信するために必要な時間をいう。

【００２６】

同期信号Ｓ１の周波数が１０．４ＭＨｚになることにより、Ｉ／Ｏブリッジ１８，２８、メモリブリッジ１６，２６は、互いに、同期信号Ｓ１の１周期で、１レーンあたり２４バイトの有効データを送信することができる。

【００２７】

演算システム１１は、プロセッサ１３、１４と、主記憶装置１５と、メモリブリッジ１６と、を備えて構成される。また、演算システム２１は、プロセッサ２３、２４と、主記憶装置２５と、メモリブリッジ２６と、を備えて構成される。

【００２８】

プロセッサ１３，１４，２３，２４は、演算処理を実行するものである。主記憶装置１５，２５は、データ等を記憶するものである。メモリブリッジ１６とプロセッサ１３，１４、メモリブリッジ２６とプロセッサ２３，２４とは、それぞれ、フロントサイドバス（ＦＳＢ）を介して接続され、クロック信号ＣＬＫに同期して動作する。

【００２９】

メモリブリッジ１６，２６は、Ｉ／Ｏブリッジ１８，２８とデータの送受信を行うものである。また、メモリブリッジ１６，２６は、通信エラー信号Ｓ２を相互に送受信する。この通信エラー信号Ｓ２は、メモリブリッジ１６，２６との間で、通信エラーの情報を共有して、連携してエラー処理を行うための信号であり、オープン・ドレイン信号として送受信される。尚、メモリブリッジ１６，２６の詳細な構成については、後述する。

【００３０】

Ｉ／Ｏシステム１２は、Ｉ／Ｏデバイス１７と、Ｉ／Ｏブリッジ１８と、コンフィギュレーションレジスタ１９と、を備えて構成される。Ｉ／Ｏシステム２２は、Ｉ／Ｏデバイス２７と、Ｉ／Ｏブリッジ２８と、コンフィギュレーションレジスタ２９と、を備えて構成される。

【００３１】

Ｉ／Ｏデバイス１７，２７は、それぞれ、Ｉ／Ｏブリッジ１８，２８との間でデータの送受信を行うものである。

【００３２】

Ｉ／Ｏブリッジ１８、２８は、それぞれ、Ｉ／Ｏデバイス１７，２７との間、又はメモリブリッジ１６，２６との間で、シリアル伝送を行うものである。

Ｉ／Ｏブリッジ１８、２８とメモリブリッジ１６，２６とは、PCI-Expressインタフェースのｘ８のリンクＬ１によって接続される。

【００３３】

即ち、演算システム１１，２１のそれぞれのメモリブリッジ１６，２６と、Ｉ／Ｏシステム１２，２２のそれぞれのＩ／Ｏブリッジ１８，２８とは、リンクＬ１を介してクロスリンク接続されている。即ち、メモリブリッジ１６は、Ｉ／Ｏシステム１２，２２に接続され、演算システム２１のメモリブリッジ２６は、Ｉ／Ｏシステム１２，２２に接続されている。

【００３４】

このようにクロスリンク接続されることにより、演算システム１１，２１は、それぞれ、Ｉ／Ｏシステム１２，２２に対して通信可能であり、Ｉ／Ｏシステム１２，２２は、それぞれ、演算システム１１，２１に対しても通信可能になる。

【００３５】

尚、PCI-Expressインタフェースでは、層単位のアップグレードを可能とするため、機能が階層化されている。そして、プロトコルが各階層毎に定義されている。

【００３６】

PCI-Expressインタフェースでは、図２（ａ）に示すように、トランザクション層において、データにヘッダが付与されてトランザクション層パケットが生成

される。

【００３７】

また、図２（ｂ）に示すように、データリンク層において、トランザクション層パケットにシーケンス番号とＣＲＣ（Cyclic Redundancy Check）とのステータス情報が付加されてデータリンク層パケット（ＤＬＬＰ：Data Link Layer Packet）が生成される。

【００３８】

そして、図２（ｃ）に示すように、物理層において、データリンク層パケットにフレームデータが付加される。そして、このパケットが送受信される。

【００３９】

Ｉ／Ｏブリッジ１８、２８は、このようなPCI-Expressインタフェースに従ってデータを送受信するためのインタフェース回路部（図示せず）を備える。

【００４０】

コンフィギュレーションレジスタ１９，２９は、分周器３１から供給された同期信号Ｓ１の１周期中において、Ｉ／Ｏブリッジ１８，２８が送信するアップストリームのパケットのパケット長、データ数を制限するためのデータを保持するものである。

【００４１】

このようにパケット長、データ数に制限を設けたのは、送信されるパケットが、送信経路長、クロックのドリフトの影響を受けないようにするためである。具体的には、送信する各パケットの最大パケット長は、１９２バイトとされる。

【００４２】

Ｉ／Ｏブリッジ１８、２８は、この制限に従い、同期信号Ｓ１の立ち上がりタイミングで、それぞれ、メモリブリッジ１６、２６に、同一のパケットを同時に送信する。

【００４３】

Ｉ／Ｏブリッジ１８，２８は、複数の小さなパケットを送信する場合、同期信号Ｓ１の１周期において送信可能な最大データ数を越えないように、送信制御を行い、１つのパケットが、周波数１０．４ＭＨｚの同期信号Ｓ１の１周期を越え

ないようにする。

【００４４】

尚、このコンフィギュレーションレジスタ１９の値をＢＩＯＳ（Basic Input/Output System）を用いて変更することができる。サブシステム１は、このようなＢＩＯＳを記憶する不揮発性メモリ（図示せず）を備える。

【００４５】

このように構成されたコンピュータシステムは、システム間の通信内容の比較により障害診断を行う。また、このコンピュータシステムは、特定のシステムでの故障が判断されると該当するシステムをマスクして、残りのシステムによって実行中の処理を継続する。

【００４６】

次に、メモリブリッジ１６，２６の構成について説明する。尚、メモリブリッジ２６はメモリブリッジ１６と同様に構成されたものであり、ここでは、メモリブリッジ１６の構成についてのみ説明する。

【００４７】

メモリブリッジ１６は、図３に示すように、インタフェース回路部４０と、同期化用バッファ５０と、内部回路部６０と、からなる。

【００４８】

インタフェース回路部４０は、PCI-Expressインタフェースに対応して設けられたものであり、データリンク／物理層４１と、トランザクション層４２と、に区分される。

【００４９】

また、データリンク／物理層４１は、物理層４３－１～４３－ｎと、データリンク層（ＲＸ）４４と、データリンク層（ＴＸ）４５と、に区分され、トランザクション層４２は、通信エラー処理部４６と、トランザクション層４７と、に区分される。

【００５０】

このデータリンク／物理層４１、トランザクション層４２、内部回路部６０は、互いに異なるクロック信号に同期して動作する。

【００５１】

物理層４３－１～４３－ｎは、同期信号Ｓ２の１周期中において、図２（ｃ）に示すようなパケットを送受信するための層である。インタフェース回路部４０は、送受信するパケットを保持するためのエラスティックバッファ（ＥＢ）を備える。尚、インタフェース回路部４０は、物理層４３－１～４３－ｎにおいて、通信エラーを検出するとエラー情報を出力する。

【００５２】

データリンク層（ＲＸ）４４は、図２（ｃ）に示すパケットからデータリンク層パケットを取り出すための層である。

【００５３】

データリンク層（ＴＸ）４５は、通信エラー処理部４６が出力したACK/NACK/flow制御信号を受け取るための層である。

【００５４】

通信エラー処理部４６は、通信エラーの処理を行うものである。

従来のPCI-Expressでは、データリンク層（ＲＸ）４４が、ステータス情報のうち、いくつかのエラー信号を、トランザクション層４７に、また、ACK/NACK/flow制御信号をデータリンク層（ＴＸ）４５に直接供給する。

【００５５】

しかし、本実施の形態では、通信エラー処理部４６が備えられ、通信エラー処理部４６がこれらのステータス情報を、データリンク層（ＲＸ）４４において取得する。そして、通信エラー処理部４６が取得したステータス情報をトランザクション層４７、データリンク層（ＴＸ）４５に供給するように構成されている。

【００５６】

通信エラー処理部４６は、データリンク層パケットに付加されたＣＲＣをチェックして、通信エラーを検出し、エラー情報を出力する。

【００５７】

通信エラー処理部４６は、物理層４３－１～４３－ｎまたはデータリンク層（ＲＸ）４４において、通信エラーが検出されなければ、トランザクション層４７とデータリンク層（ＴＸ）４５とに、データ、ステータス情報をそのまま出力す

る。インタフェース回路部４０は、受信したデータに通信エラーがなければ、ステータス情報に従って、データを送信したＩ／Ｏブリッジ１８，２８に定期的にＡＣＫ信号を返送する。

【００５８】

一方、通信エラー処理部４６は、物理層４３－１～４３－ｎまたはデータリンク層（ＲＸ）４４において、通信エラーが検出されると、同期信号Ｓ１の１周期の間に受信した全パケットをロストパケットとしてキャンセルし、受信データのトランザクション層４７への出力を停止する。

【００５９】

通信エラー処理部４６は、パケットをキャンセルすると、データリンク層（ＲＸ）４４に対して、次に受信する予定のパケットのシーケンス番号を、通信エラーパケット受信前のシーケンス番号に設定するように指示する。

【００６０】

また、通信エラー処理部４６は、通信エラーを検出すると、同期信号Ｓ１の１周期の間、通信エラー信号Ｓ２をアサート（assert）、即ち、有効とする。通信エラー処理部４６は、アサートした通信エラー信号Ｓ２を、信号線を介してメモリブリッジ２６に送信する。

【００６１】

トランザクション層４７は、上位のソフトウェア層からの読み出しと書き込み要求を受け付けるとともにデータリンク層（ＲＸ）４４、データリンク層（ＴＸ）４５に対してパケットの転送を要求するための層である。

【００６２】

同期化用バッファ５０は、トランザクション層４７と内部回路部６０との間で、データの受け渡しをするためのバッファであり、トランザクション層４７から出力されたデータを保持する。

【００６３】

内部回路部６０は、同期化用バッファ５０が保持しているデータを同期信号Ｓ１に同期したタイミングで取得し、取得したデータをプロセッサ１３，１４、主記憶装置１５に送出するための回路である。

【００６４】

尚、Ｉ／Ｏブリッジ１８，２８がメモリブリッジ１６，２６にシリアルデータを送信する場合、Ｉ／Ｏブリッジ１８，２８が送信インタフェースとなり、メモリブリッジ１６，２６のインタフェース回路部が受信インタフェース部となる。

【００６５】

また、前述のようにメモリブリッジ１６，２６がＩ／Ｏブリッジ１８，２８にシリアルデータを送信することも可能であり、この場合、メモリブリッジ１６，２６が送信インタフェースとなり、Ｉ／Ｏブリッジ１８，２８のインタフェース回路部が受信インタフェース部となる。

【００６６】

次に本実施の形態に係るコンピュータシステムの動作を説明する。

尚、ここでは、Ｉ／Ｏブリッジ１８が、メモリブリッジ１６，２６にシリアルデータを送信する場合について説明する。

【００６７】

Ｉ／Ｏブリッジ１８は、Ｉ／Ｏデバイス１７からデータが供給されると、Ｉ／Ｏブリッジ１８は、図２（ａ）に示すように、トランザクション層において、シリアルデータにヘッダを付与し、トランザクション層パケットを生成する。

【００６８】

Ｉ／Ｏブリッジ１８は、図２（ｂ）に示すように、データリンク層において、生成したトランザクション層パケットに、シーケンス番号とＣＲＣとのステータス情報を付加し、データリンク層パケットを生成する。

【００６９】

そして、Ｉ／Ｏブリッジ１８は、図２（ｃ）に示すように、物理層において、生成したデータリンク層パケットに、フレームデータを付加する。そして、Ｉ／Ｏブリッジ１８は、図２（ｃ）に示すパケットを、リンクＬ１を介してメモリブリッジ１６，２６に送信する。

【００７０】

メモリブリッジ１６のインタフェース回路部４０は、物理層４３－１～４３－ｎにおいて、このデータを受信する。

インタフェース回路部４０は、図４（ａ）に示すように、物理層４３－１～４３－ｎにおいて、同期信号Ｓ１の１周期中において受信した全パケットを、エラスティックバッファに、一旦、格納してから、データリンク層（ＲＸ）４４に出力する。

【００７１】

インタフェース回路部４０は、データリンク層（ＲＸ）４４において、図２（ｃ）に示すパケットからデータリンク層パケットを取り出す。また、インタフェース回路部４０は、図２（ｂ）に示すデータリンク層パケットに含まれているＣＲＣに基づいてエラーの検出を行う。

【００７２】

通信エラー処理部４６は、図４（ｃ）に示すように、同期信号Ｓ１の各周期中において受信したパケットを、それぞれ、次の同期信号Ｓ１の立ち上がりに同期して、取得する。

【００７３】

受信したパケットにエラーが検出されなければ、通信エラー処理部４６は、図４（ｄ）に示すように、通信エラー信号Ｓ２をハイ（Ｈ）レベルにしてディアサート、即ち、無効とする。従って、受信されたパケットは有効となる。

【００７４】

そして、通信エラー処理部４６は、図４（ｅ）に示すように、各パケットを、次の同期信号Ｓ１の立ち上がりに同期してトランザクション層に送出する。

【００７５】

インタフェース回路部４０は、トランザクション層において、データリンク層パケットから、トランザクション層パケットを取り出し、さらに、データと取り出して、図４（ｆ）に示すように同期化用バッファ５０に出力する。

【００７６】

内部回路部６０は、同期信号Ｓ１の立ち上がりに同期して、同期化用バッファ５０からデータを取得し、取得したデータをプロセッサ１３，１４、主記憶装置１５に出力する。

【００７７】

Ｉ／Ｏブリッジ１８からメモリブリッジ１６，２６にデータを送信する場合、Ｉ／Ｏブリッジ１８とメモリブリッジ１６との間と、Ｉ／Ｏブリッジ１８とメモリブリッジ２６との間で、リンクＬ１の配線長に差異がほぼなければ、図５（ｂ），（ｃ）に示すように、メモリブリッジ１６，２６は、ほぼ同時にデータを受信する。

【００７８】

しかし、Ｉ／Ｏブリッジ２８とメモリブリッジ２６との間のリンクＬ１が、Ｉ／Ｏブリッジ１８とメモリブリッジ１６との間のリンクＬ１よりも長いと、図５（ｄ），（ｅ）に示すように、メモリブリッジ１６，２６がデータを受信するタイミングに差異が生じる。

【００７９】

但し、差異が生じても、同じ同期信号Ｓ１の周期内であれば、演算システム１１，２１は、クロック信号ＣＬＫに同期して同じ処理を実行する。

もし、メモリブリッジ２６が、同期信号Ｓ１の１周期目と２周期目とでデータを受信するようになるのであれば、Ｉ／Ｏブリッジ１８は、１パケットにおけるデータの長さを短くするように、ＢＩＯＳを用いてコンフィギュレーションレジスタ１９が格納しているデータを変更する。

【００８０】

次に、図６（ａ）、（ｂ）に示すように、メモリブリッジ１６のインタフェース回路部４０が、同期信号Ｓ１の第２クロックサイクルにおいて、受信したパケットの通信エラーを検出するものとする。

【００８１】

この場合、通信エラー処理部４６は、図６（ｃ）、（ｅ）に示すように、第３クロックサイクルにおいて、データリンク層パケット(DLLP)をその一部に含んでいたとしても、すべてのパケットをキャンセル扱いとする。

【００８２】

そして、通信エラー処理部４６は、第３クロックサイクル以降のパケットもキャンセルする。通信エラー処理部４６は、第２クロックサイクルでキャンセルしたパケットが再送されるまで、すべてのパケットの受信をキャンセルする。

【００８３】

通信エラー処理部４６は、通信エラーが検出されると、データリンク層（ＲＸ）４４において、管理されているパケットのシーケンス番号をエラー発生前の番号に設定し直す。

【００８４】

また、通信エラーが検出されると、通信エラー処理部４６は、図６（ｄ）に示すように、通信エラー信号Ｓ２をロー（Ｌ）レベルとしてアサート、即ち、有効とする。同期信号Ｓ１の第４クロックサイクルになると、パケットは受信されないため、通信エラー処理部４６を、ディアサートする。

【００８５】

通信エラー処理部４６は、データの送信元であるＩ／Ｏブリッジ１８に、データの再送を要求する。Ｉ／Ｏブリッジ１８は、メモリブリッジ１６から再送要求を受信した場合、再送要求のあったパケットを再送する。また、メモリブリッジ１６からのACK信号が返送されずに、所定期間が経過した場合にも、送信が確認されていないパケットの再送信を行う。

【００８６】

続いて、図７（ｂ）に示すように、メモリブリッジ１６が、同期信号Ｓ１の第６クロックサイクルにおいて、再送要求に応答してシーケンス番号２のパケットを受信し、通信エラーがなければ、インタフェース回路部４０は、図６（ｃ）～（ｅ）に示すように、キャンセルされた第３クロックサイクル以降のパケットを、そのまま、受信する。

【００８７】

次に、図８（ａ）～（ｃ）に示すように、メモリブリッジ１６が受信したデータに通信エラーは検出されなくても、メモリブリッジ２６が通信エラーを検出すると、メモリブリッジ２６は、図８（ｄ）に示すように、メモリブリッジ１６にローレベルの通信エラー信号Ｓ２を出力する。

【００８８】

メモリブリッジ２６は、同期信号Ｓ１の第３クロックサイクルにおいて、通信エラー信号Ｓ２がアサートされると、通信エラー処理部４６は、図８（ｅ）に示

すように、同期信号Ｓ１の第３クロックサイクルにおいて保持しているシーケンス番号２のパケットをキャンセルする。

【００８９】

通信エラー処理部４６は、同期信号Ｓ１の第４クロックサイクル以降、トランザクション層４７へのパケットの引き渡しを停止する。

そして、通信エラー処理部４６は、データリンク層（ＲＸ）４４において、管理する次回受信予定パケットのシーケンス番号を、パケットキャンセル前の値に再設定させる。

【００９０】

図９（ａ），（ｂ）に示すように、メモリブリッジ１６が、通信エラーを有するパケットに続いて、データリンク層パケットのみで構成されるパケットを受信した場合、通信エラー処理部４６は、図６（ｃ），（ｅ）に示すように、まず、通信エラーを有するパケットをキャンセルする。

【００９１】

通信エラー処理部４６は、通信エラーを有するパケットをキャンセルすると、図９（ｄ）に示すように、通信エラー信号Ｓ２をアサートして、シーケンス２以降のパケット列の再送信を要求する。しかし、通信エラー処理部４６は、通信エラーを有するパケットをキャンセルしても、図６（ｃ），（ｅ）に示すように、同期信号Ｓ１の第３クロックサイクルにおいて受信されたデータリンク層パケットをキャンセルしない。これは、データリンク層パケットには、シーケンス番号がないため、シーケンス番号エラーが発生しないためである。

【００９２】

エラー処理部４６は、データリンク層パケットをキャンセルしなくても、再送されたパケットのシーケンス番号から、順序を特定することができ、メモリブリッジ１６は、問題なく再送信されたパケットを受信することができる。

【００９３】

以上説明したように、本実施の形態によれば、メモリブリッジ１６のインタフェース回路部４０が通信エラーを検出した場合、通信エラー処理部４６が受信したパケットをキャンセルする。そして、通信エラー処理部４６は、アサートした

通信エラー信号Ｓ２をメモリブリッジ２６に出力し、キャンセルしたパケットの再送をパケット送信元に要求するようにした。

【００９４】

従って、通信エラーが生じた場合でも、メモリブリッジ１６，２６の通信エラー処理部が連携してパケット送信元に、パケットの再送信をするようになるため、受信データの同期ずれを回避することができる。そして、演算システム１１，２１は、同一のデータ処理を同期して行うことができる。

【００９５】

また、Ｉ／Ｏブリッジ１８が、リンクＬ１の配線長に差異が生じても、影響を受けないように、送信するパケットのパケット長、データ数に制限が設けるようにした。

【００９６】

従って、フォールトトレラントコンピュータシステムを構成するための回路基盤設計および筐体設計が容易になる。

【００９７】

尚、本発明を実施するにあたっては、種々の形態が考えられ、上記実施の形態に限られるものではない。

例えば、上記実施の形態では、メモリブリッジ１６，２６、Ｉ／Ｏブリッジ１８，２８に、インタフェース回路部を備えるようにした。しかし、図１０に示すように、演算システム１１，２１において、メモリブリッジ１６，２６とは別に、それぞれ、送受信ブリッジ７１，７２を備え、また、Ｉ／Ｏシステム１２，２２とは別に、それぞれ、送受信ブリッジ８１，８２を備えることができる。

【００９８】

この場合、送受信ブリッジ７１，７２，８１，８２は、それぞれ、通信エラー処理部を備える。送受信ブリッジ７１，７２は、それぞれ、メモリブリッジ１６，２６に接続され、送受信ブリッジ８１，８２は、それぞれ、Ｉ／Ｏブリッジ１８，２８に接続される。

【００９９】

そして、送受信ブリッジ７１，７２と、送受信ブリッジ８１，８２とは、互い

にロックステップ方式に従って同期化されてデータの送受信を行う。尚、送受信ブリッジ７１，７２は、既存のメモリブリッジ１６，２６とは、１セットの通信リンクで接続される。接続は、既存のメモリブリッジのサポートする高速シリアルリンクによって行われる。この場合、メモリブリッジ１６と送受信ブリッジ７１、メモリブリッジ２６と送受信ブリッジ７２との間のリンクの配線長は、受信タイミングの差異による通信エラーの発生を避けるため、できるだけ、短くされる。

【０１００】

このように構成されることにより、既存のシステムチップセットコンポーネントをそのまま利用して、フォールトトレラントコンピュータシステムを構成することができる。

【０１０１】

本実施の形態では、コンピュータシステムに２つのサブシステム１，２を備え、また、サブシステム１，２に、それぞれ、２つのプロセッサ１３，１４，２３，２４を備えて、コンピュータシステムを二重化冗長構成とした。しかし、コンピュータシステムの構成は、これに限られるものではなく、三重化冗長構成、あるいはそれ以上の冗長構成とすることができる。

【０１０２】

また、本実施の形態では、高速シリアルリンクとしてPCI-Expressを例として説明した。しかし、リンクとしては、これに限られるものではなく、InfiniBand、HyperTransport等の他の高速シリアルリンクを用いることもできる。

【０１０３】

また、本実施の形態では、メモリブリッジ１６，２６と、Ｉ／Ｏブリッジ１８，２８との間で送受信するデータをシリアルデータとして説明したが、パラレルデータであっても、本実施の形態を適用することができる。

【０１０４】

【発明の効果】

以上説明したように、本発明によれば、通信エラーが生じた場合でも同一のデータ処理を同期して行うことができる。

【図面の簡単な説明】

【図１】

本発明の実施の形態に係るコンピュータシステムの構成を示すブロック図である。

【図２】

送受信されるパケットデータの構成を示す説明図である。

【図３】

図１に示すメモリブリッジの詳細な構成を示すブロック図である。

【図４】

図１に示すメモリブリッジの動作を示すタイミングチャートである。

【図５】

図１に示すメモリブリッジの動作を示すタイミングチャートである。

【図６】

図１に示すメモリブリッジの動作を示すタイミングチャートである。

【図７】

図１に示すメモリブリッジの動作を示すタイミングチャートである。

【図８】

図１に示すメモリブリッジの動作を示すタイミングチャートである。

【図９】

図１に示すメモリブリッジの動作を示すタイミングチャートである。

【図１０】

本発明の実施の形態に係るコンピュータシステムの応用例を示すブロック図である。

【符号の説明】

１，２　サブシステム

１３、１４，２３，２４　プロセッサ

１６，２６　メモリブリッジ

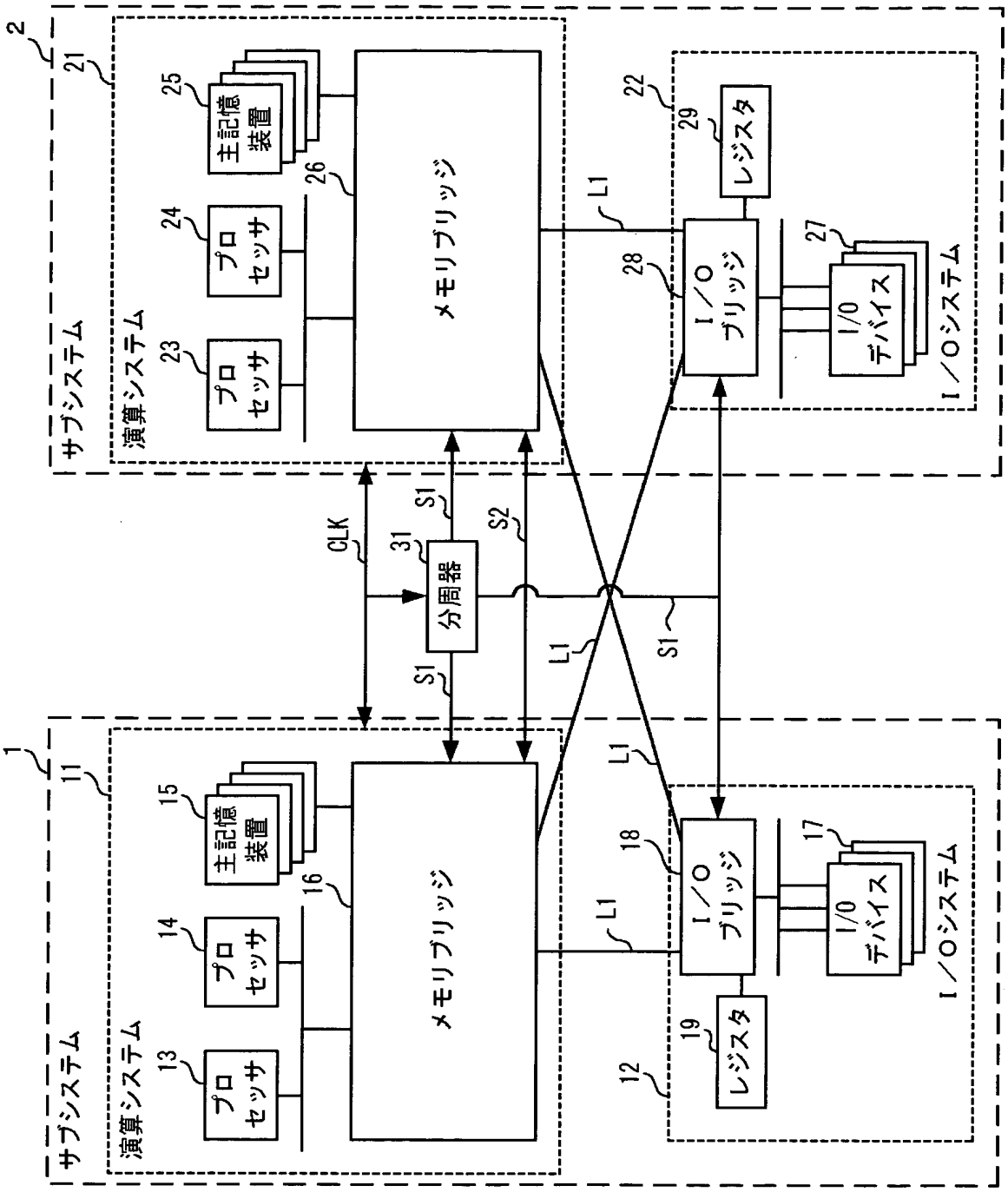
１８，２８　Ｉ／Ｏブリッジ

４６　通信エラー処理部

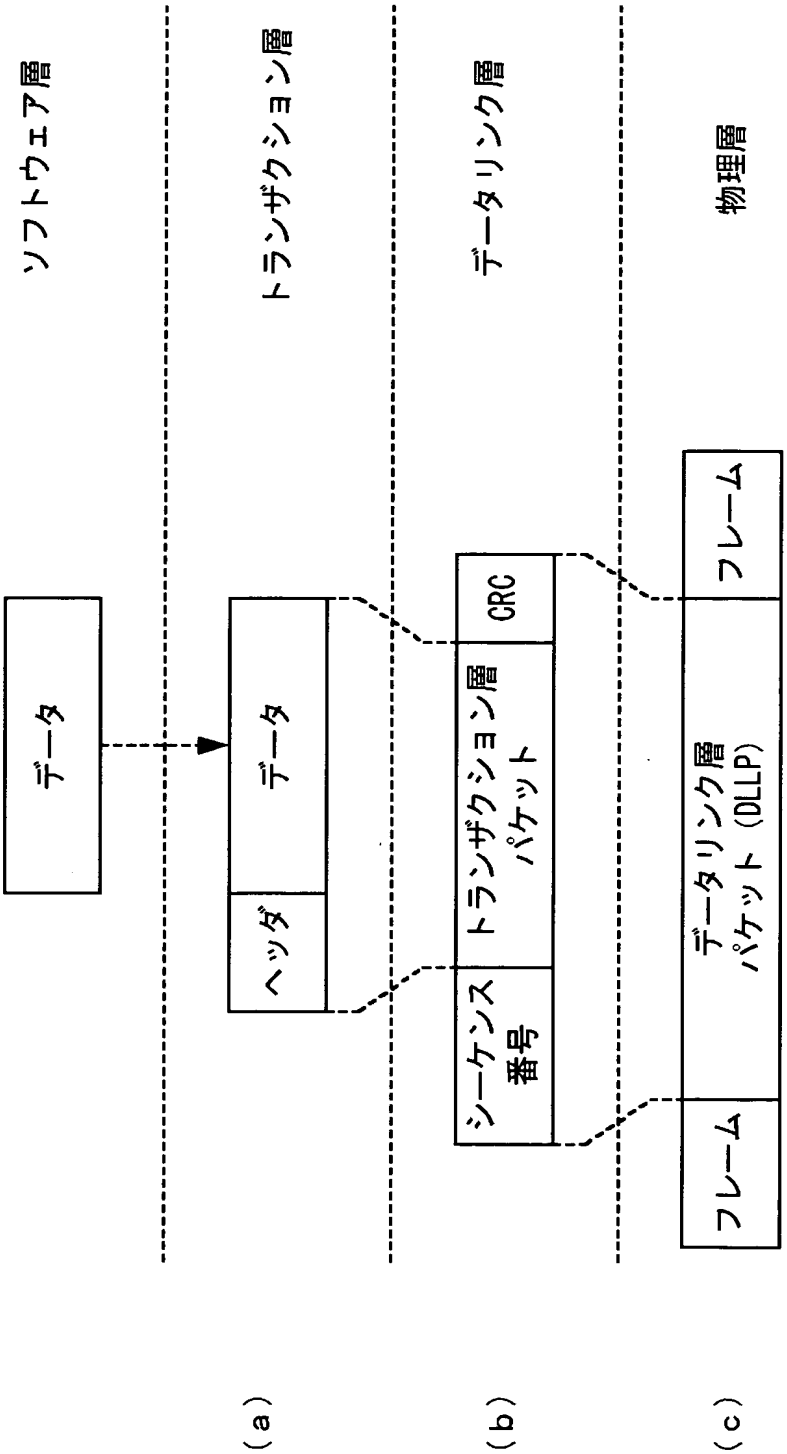
３１　分周器

【書類名】　図面

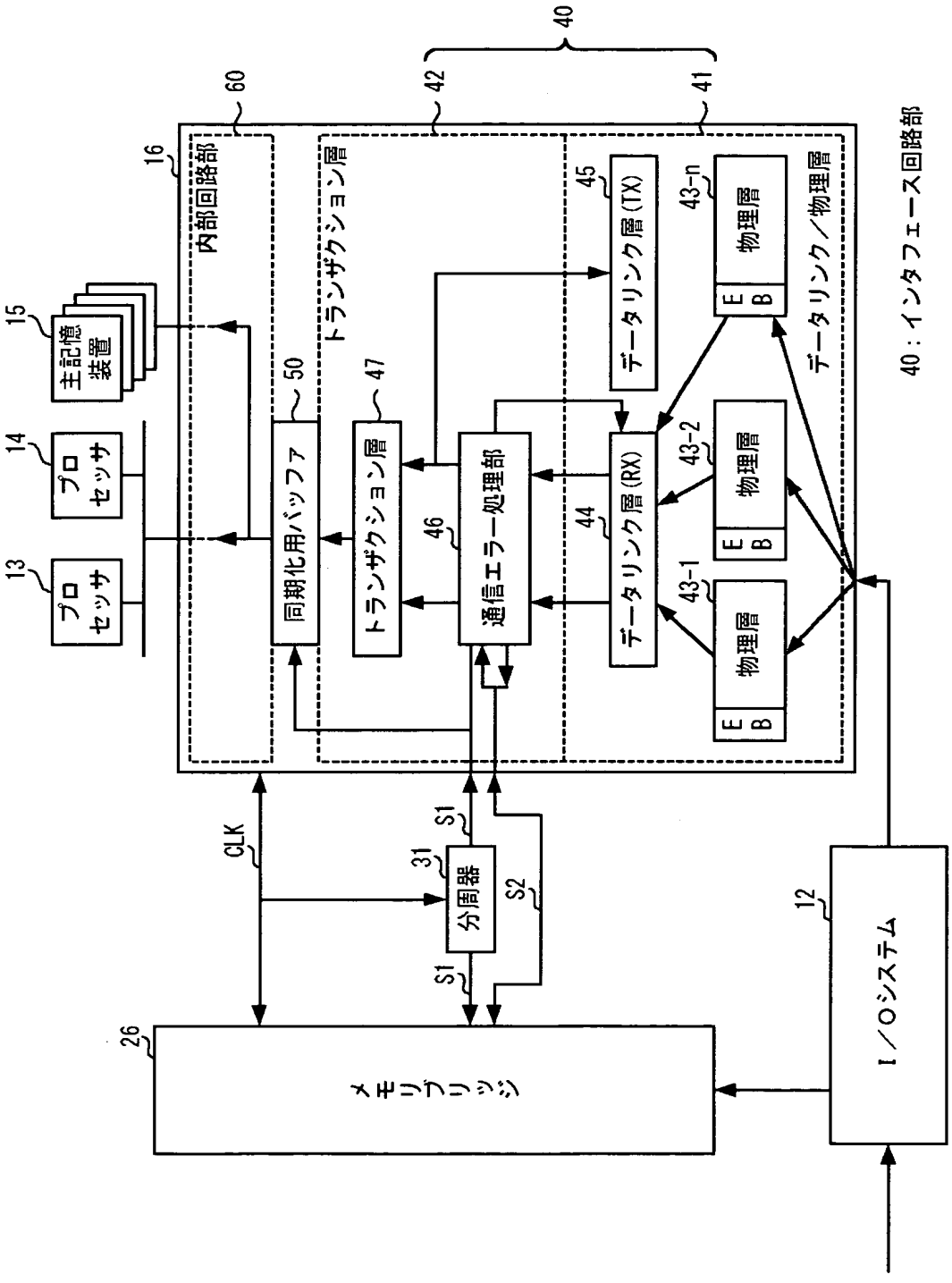
【図１】


2
21
サブシステム
演算システム
25
主記憶装置
24
プロセッサ
23
プロセッサ
26
メモリブリッジ
22
29
レジスタ
L1
28
I／Oブリッジ
27
I/Oデバイス
I／Oシステム
CLK
31
分周器
S1
S2
S1
L1
S1
1
11
演算システム
15
主記憶装置
14
プロセッサ
13
プロセッサ
16
メモリブリッジ
L1
18
I／Oブリッジ
17
I/Oデバイス
I／Oシステム
12
19
レジスタ

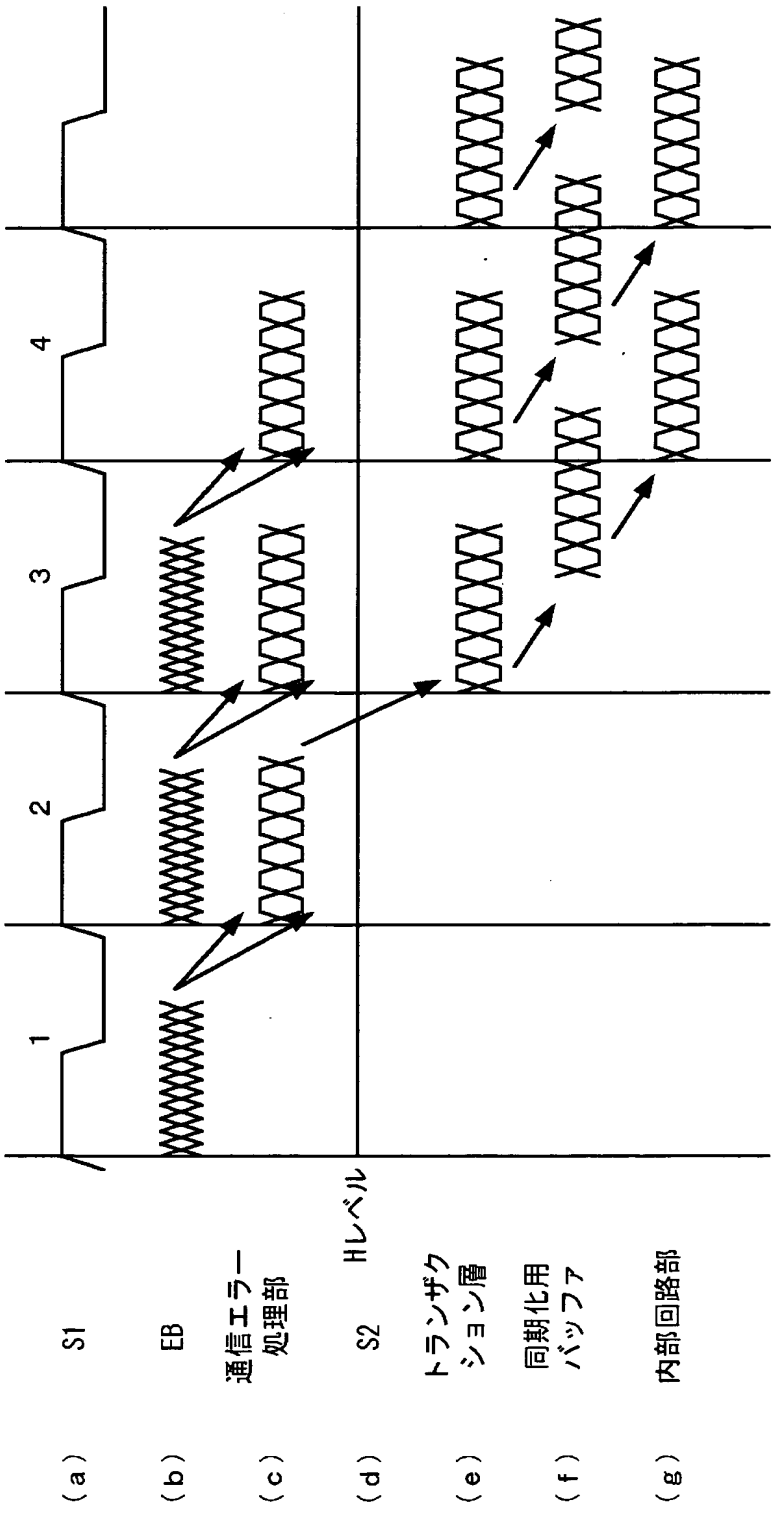
【図２】


ソフトウェア層
トランザクション層
データリンク層
物理層
データ
データ
ヘッダ
トランザクション層
パケット
CRC
シーケンス
番号
データリンク層
パケット（DLLP）
フレーム
フレーム
（a）
（b）
（c）

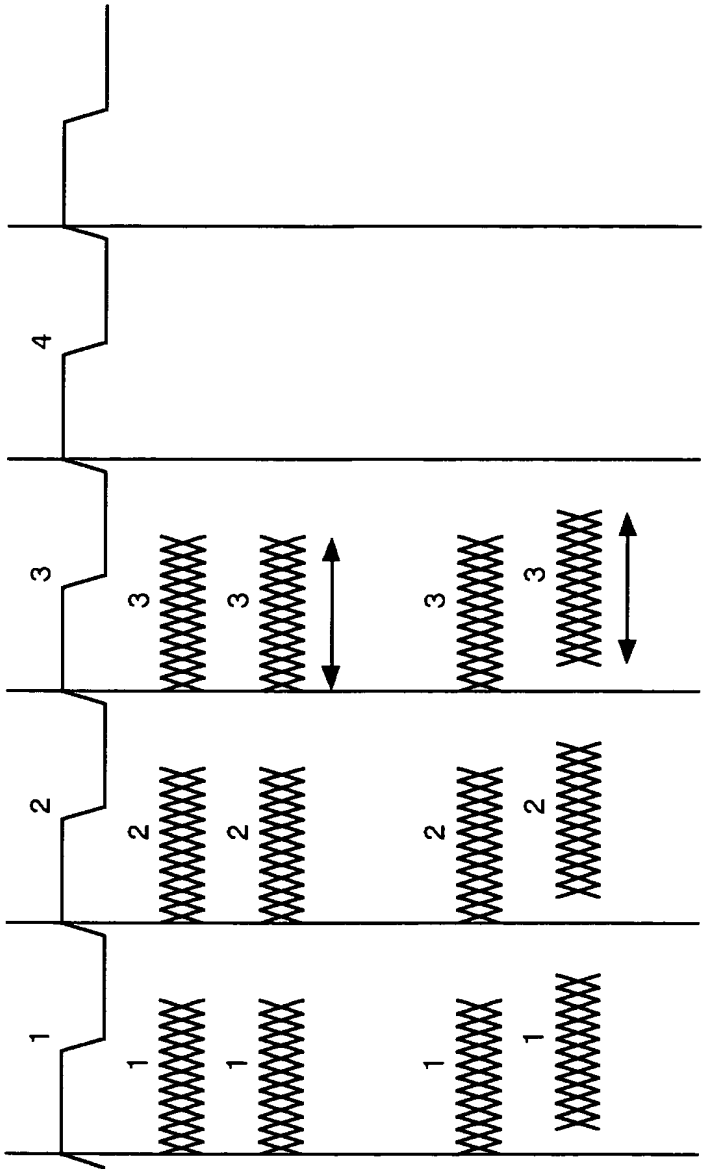
【図３】


40
60
42
41
16
内部回路部
トランザクション層
データリンク／物理層
40：インタフェース回路部
15
主記憶装置
14
プロセッサ
13
プロセッサ
50
同期化用バッファ
47
トランザクション層
46
通信エラー処理部
44
45
データリンク層(TX)
データリンク層(RX)
43-n
43-2
43-1
物理層
E B
CLK
S1
S2
31
分周器
26
メモリリンク
12
I／Oシステム

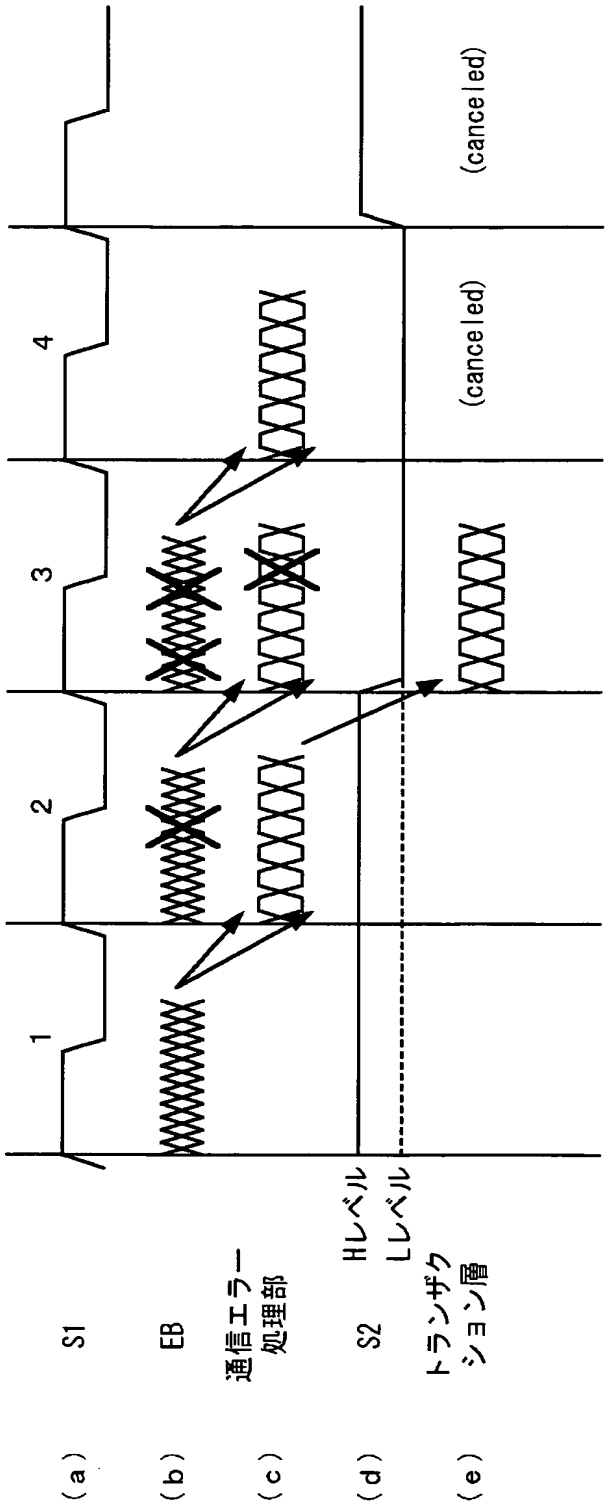
【図４】


1
2
3
4
(a) S1
(b) EB
(c) 通信エラー処理部
(d) S2 Hレベル
(e) トランザクション層
(f) 同期化用バッファ
(g) 内部回路部

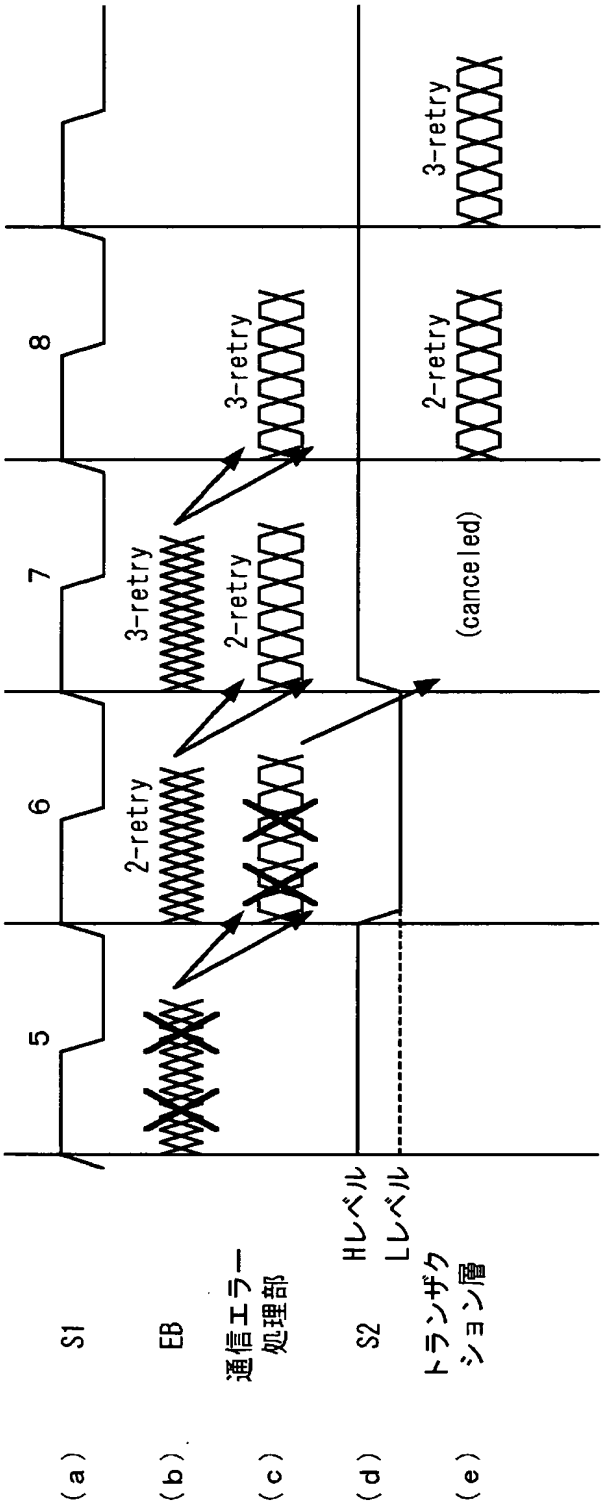
【図５】


1
2
3
4
1
2
3
(a) SI
(b) メモリ ブリッジ16
(c) メモリ ブリッジ26
(d) メモリ ブリッジ16
(e) メモリ ブリッジ26

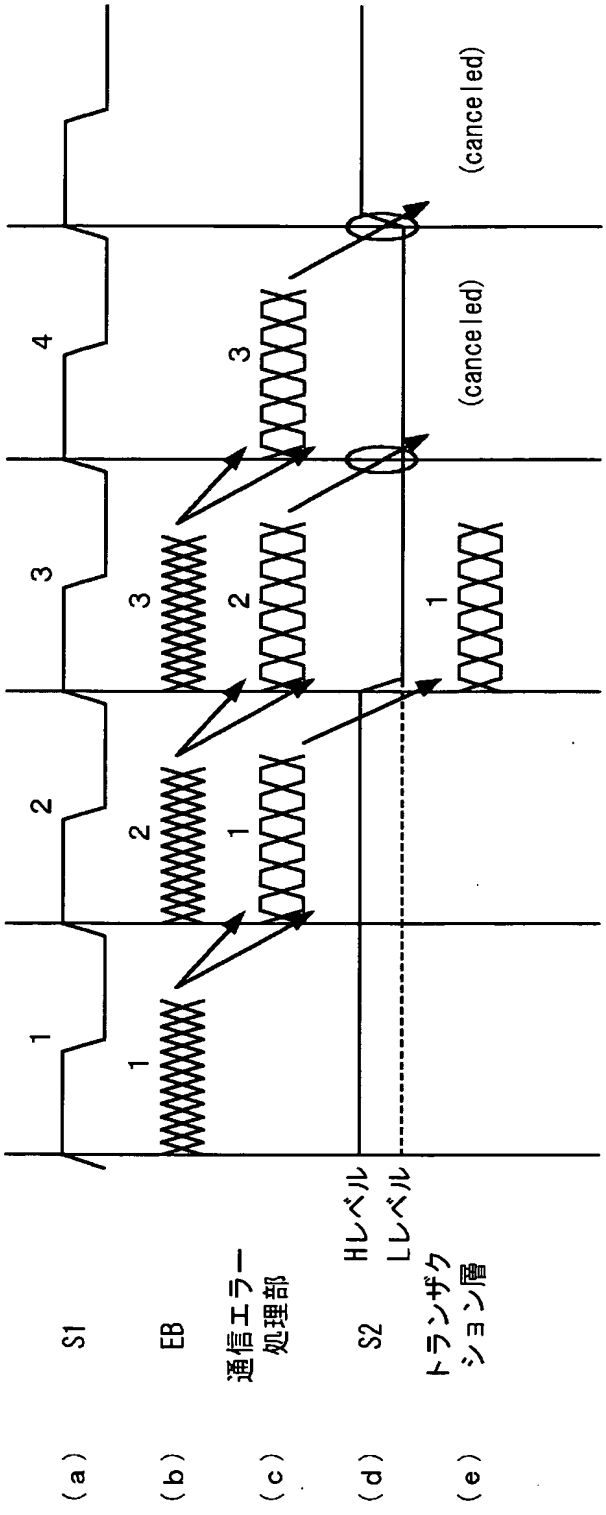
【図６】


1
2
3
4
(canceled)
(canceled)
S1
EB
通信エラー処理部
S2
Hレベル
Lレベル
トランザクション層
(a)
(b)
(c)
(d)
(e)

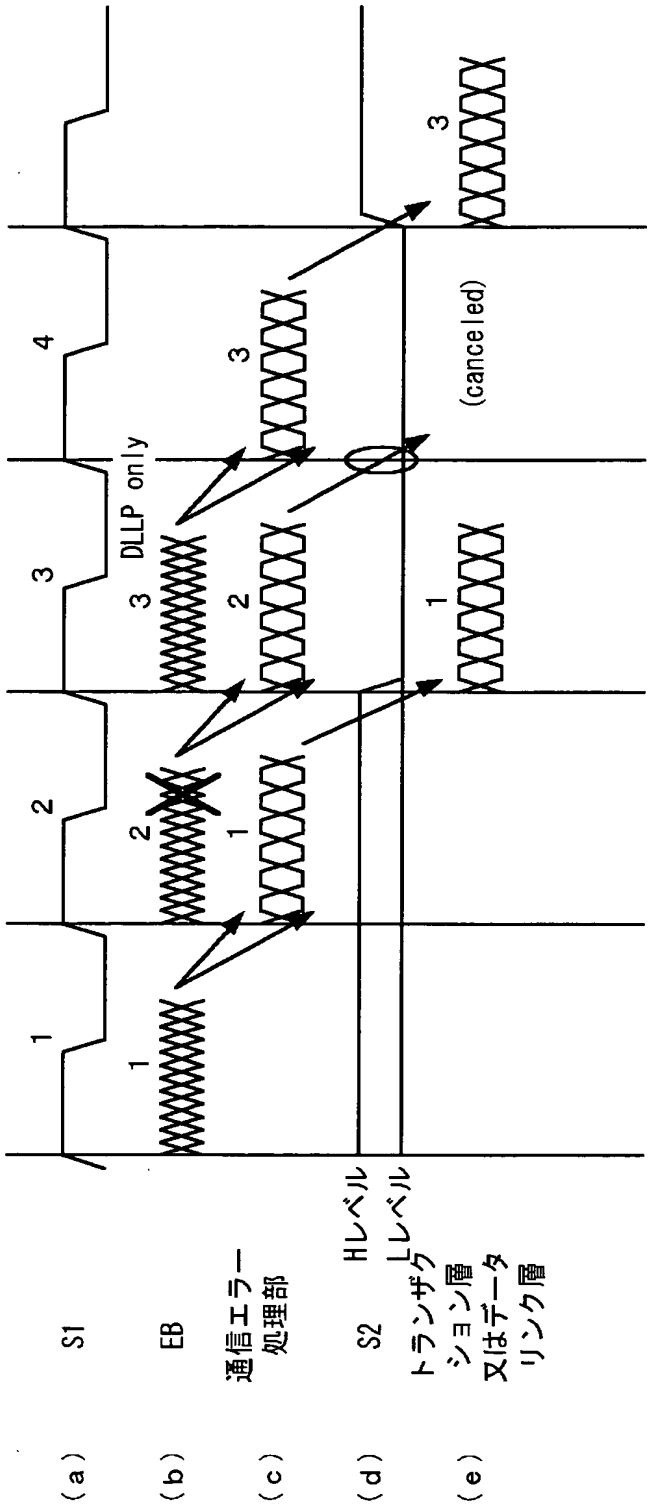
【図７】


5
6
7
8
2-retry
3-retry
2-retry
3-retry
2-retry
(canceled)
3-retry
(a)
S1
(b)
EB
(c)
通信エラー処理部
(d)
S2
Hレベル
Lレベル
(e)
トランザクション層

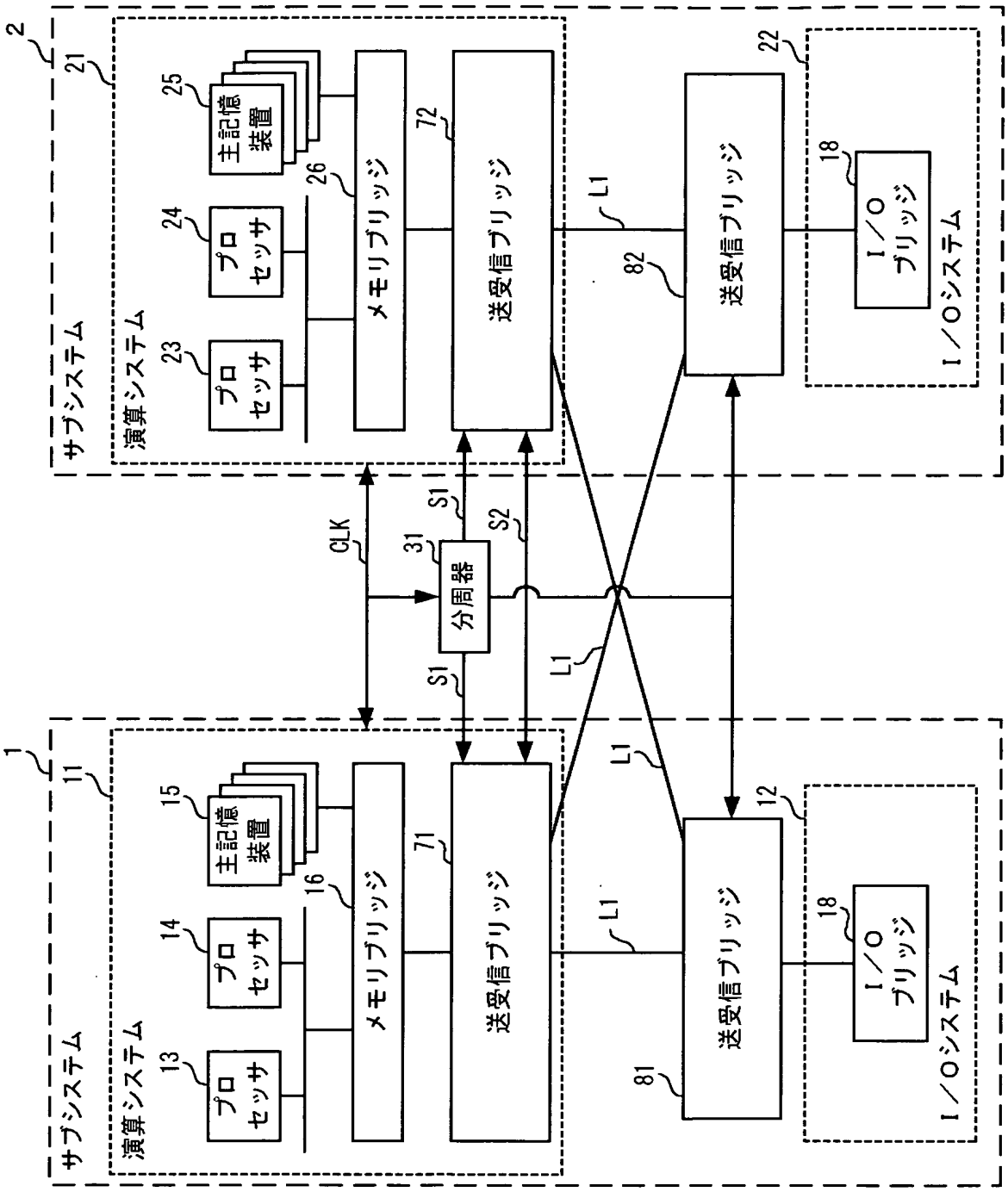
【図８】


(canceled)
(canceled)
4
3
3
2
2
1
1
3
2
1
(a)
(b)
(c)
(d)
(e)
S1
EB
通信エラー処理部
S2
Hレベル
Lレベル
トランザクション層

【図９】


(a)
S1
1
2
3
4
(b)
EB
1
2
3 DLLP only
(c)
通信エラー
処理部
1
2
3
(d)
S2
Hレベル
Lレベル
(e)
トランザクション層
又はデータリンク層
1
2
3
(canceled)

【図１０】


2
サブシステム
21
演算システム
25
主記憶装置
24
プロセッサ
23
プロセッサ
26
メモリブリッジ
72
送受信ブリッジ
L1
82
送受信ブリッジ
22
I／Oシステム
18
I／Oブリッジ
CLK
31
分周器
S1
S2
1
サブシステム
11
演算システム
15
主記憶装置
14
プロセッサ
13
プロセッサ
16
メモリブリッジ
71
送受信ブリッジ
81
送受信ブリッジ
12
I／Oシステム
18
I／Oブリッジ

【書類名】　要約書

【要約】

【課題】　通信エラーが生じた場合でも同一のデータ処理を同期して行えるようにする。

【解決手段】　クロスリンク接続されたメモリブリッジ１６，２６と、Ｉ／Ｏブリッジ１８，２８とには、PCI-Expressインタフェースに従ってデータの送受信を行うためのインタフェース回路部が、それぞれ、備えられる。また、各インタフェース回路部には、通信エラー処理部が備えられる。メモリブリッジ１６の通信エラー処理部は、Ｉ／Ｏブリッジ１８から受信したデータにエラーが発生すると、受信したデータをキャンセルしてメモリブリッジ２６に通信エラー信号を出力する。メモリブリッジ２６は、この通信エラー信号が供給されてデータの受信を停止する。そして、メモリブリッジ１６の通信エラー処理部は、Ｉ／Ｏブリッジ１８にデータの再送信を要求する。

【選択図】　図１

特願２００３－１１５６２１

出　願　人　履　歴　情　報

識別番号　［０００００４２３７］

１．変更年月日　１９９０年　８月２９日
［変更理由］　新規登録
住　所　東京都港区芝五丁目７番１号
氏　名　日本電気株式会社